

ABSTRACT OF THE DISCLOSURE

A system, method, and processor readable medium for normalizing documents using extensible markup language (XML). The system may determine a type of object repository storing at least one object. The object may include metadata. The system may then identify the object stored in the object repository. At least one portion of the one object may be extracted from the repository, wherein the portion is extracted in extensible markup language (XML) format. Preferably, some of the metadata is preserved. The metadata preserved may include at least one of author, title, subject, date created, date modified, list of modifiers, and link list information. The portion may then be transmitted to a processor. The processor may perform one or more processes on the portion. A mapping may be performed that maps at least one field in the object with a field designation identifier. The processor may include at least one of a full-text engine, a metrics engine, and a taxonomy engine.